Predição de Mapas de Contatos de Proteínas

Aluno: Lucas G. Farris
Orientador: Jacques Wainer
Coorientador: Guilherme P. Telles

Universidade Estadual de Campinas
Instituto de Computação

Campinas, 24 de Setembro de 2015

Roteiro

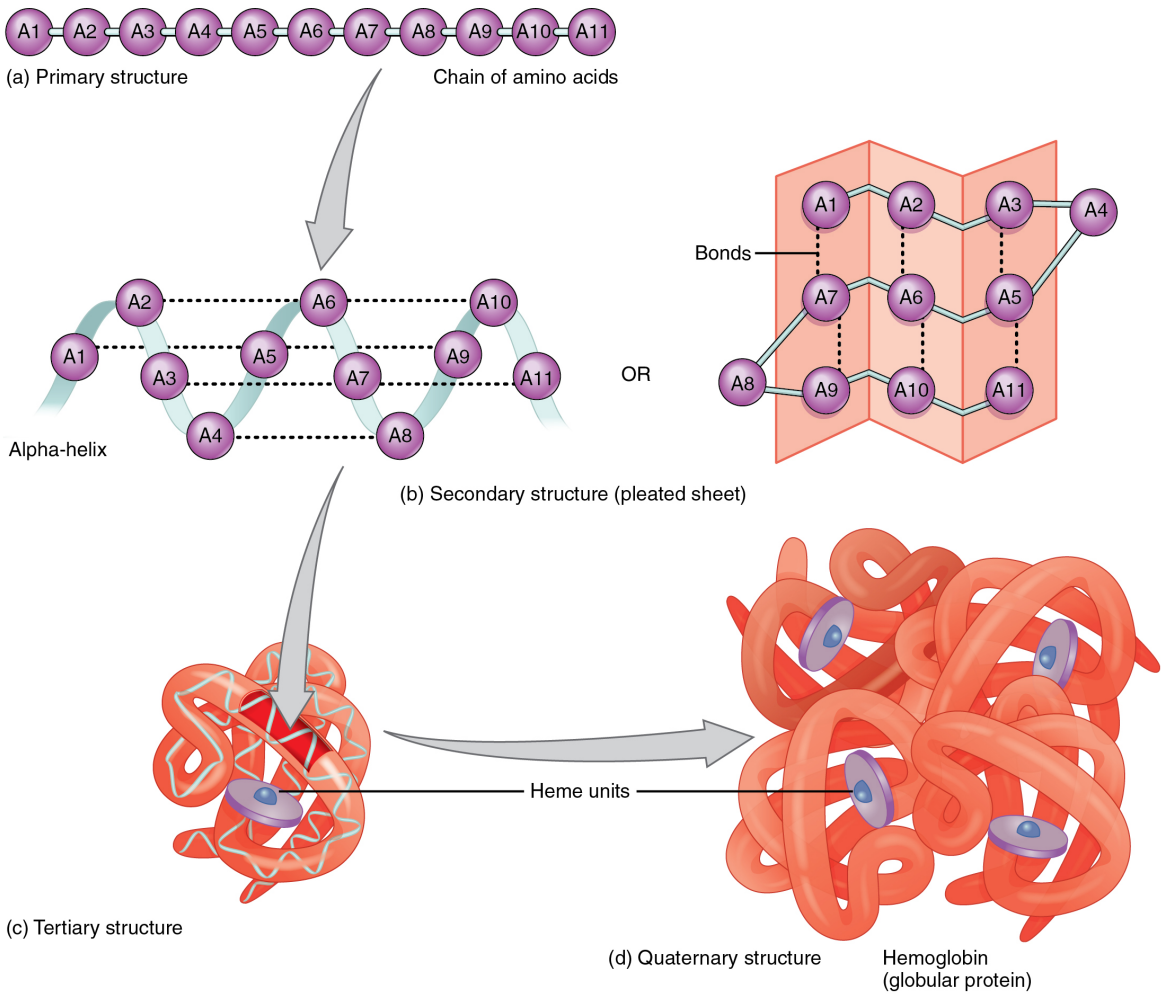
Motivação

- Proteínas são parte fundamental dos organismos vivos. Elas compõe nossos tecidos, bem como enzimas e anticorpos.
- Descobrir suas estruturas tridimensionais ajuda a descobrir mais sobre elas.
- Sequenciamento é barato, mas cristalografia é cara.

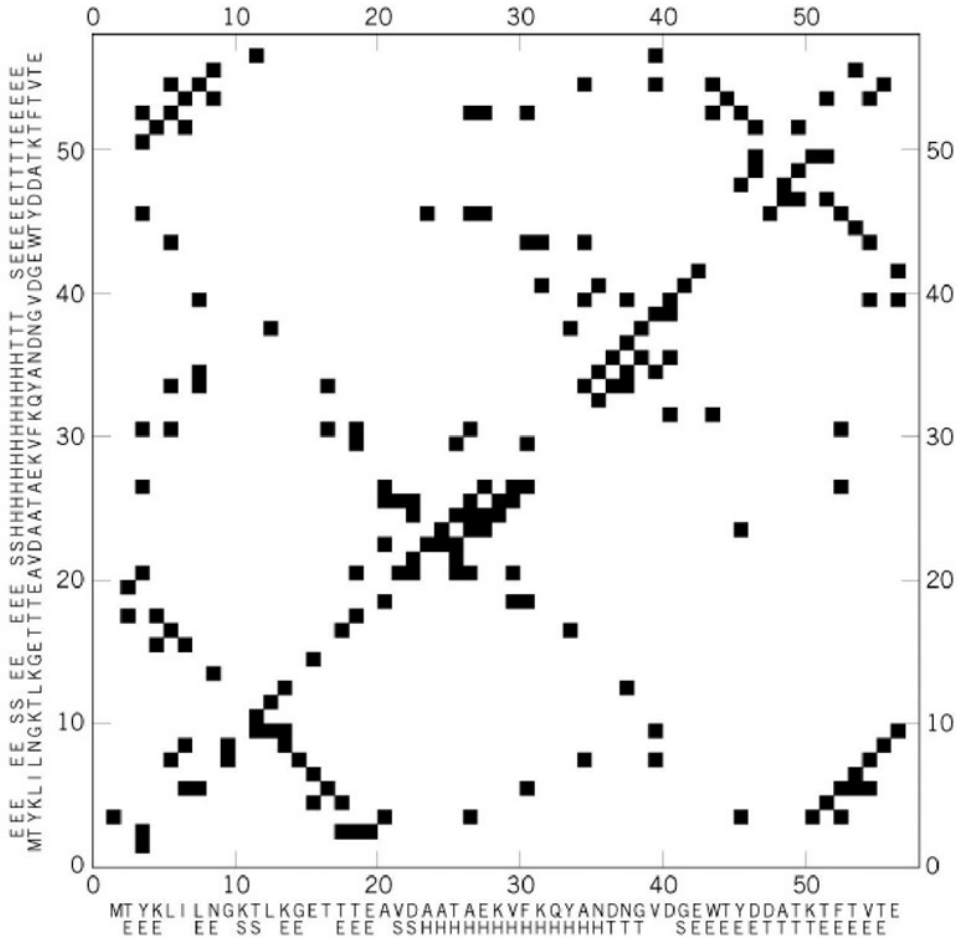
Conceitos Biológicos

- Resíduos (Aminoácidos)
- Estruturas primária, secundária, terciária e quaternária
- Contatos
- Short, Medium e Long-Range
- Mapas de Contatos Resíduo-Resíduo

Estruturas


A1 A2 A3 A4 A5 A6 A7 A8 A9 A10 A11
(a) Primary structure
Chain of amino acids
Alpha-helix
Bonds
OR
(b) Secondary structure (pleated sheet)
Heme units
(c) Tertiary structure
(d) Quaternary structure
Hemoglobin
(globular protein)

Mapa de Contato



Revisão Bibliográfica

- Redes Neurais (1990 até 2001)[FC99][Far+01]
- *Support Vector Machines* (2007) [CB07]
- *Random Forests* (2011) [YF11]
- *Boosting* (2012)[EC12]
- *Deep Learning* (desde 2012)[DNB12]

Features

- Globais: contagem de aminoácidos.
- Contatos: tipo, propensão, acessibilidade de solvente e pontos isoelétricos.
- Janelas dos contatos: estruturas secundárias, acessibilidade e médias de pontos isoelétricos.
- Entre os contatos: tamanho do intervalo, distribuição de estruturas secundárias, tripeptídeos e informações do resíduo central.

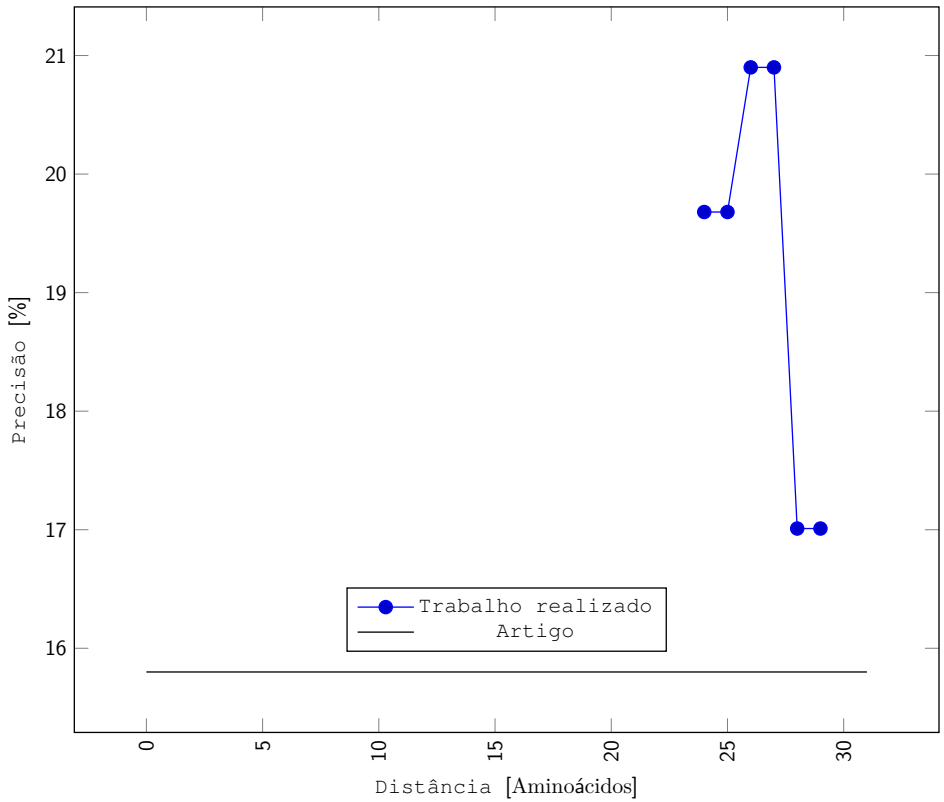
Trabalho Desenvolvido

- Conjunto não redundante de proteínas.
- Algoritmo para extração de *features*.
- Implementação em **Python** de *Random Forests*.
- Acurácia *long-range* comparável a resultados de artigo.

Resultados Parciais I

- Treinamento com 1490 proteínas, validação com 329
- Floresta com 1500 árvores, profundidade máxima 21
- Balanceamento artificial com o dobro de *negative samples*
- Treinamos um modelo por distância

Resultados Parciais II



Próximos Passos

- Explorar conjuntos de dados maiores.
- *Features* da literatura podem ser melhoradas.
- Encontrar modelos adequados para os dados.

Referências I

Jianlin Cheng and Pierre Baldi. 'Improved residue contact prediction using support vector machines and a large feature set'. In: *BMC Bioinformatics* 8.113 (2007).

Pietro Di Lena, Ken Nagata and Pierre Baldi. 'Deep architectures for protein contact map prediction'. In: *Bioinformatics (Oxford, England)* 28.19 (2012), pp. 2449–2457.

Referências II

Jesse Eickholt and Jianlin Cheng. 'Predicting protein residue-residue contacts using deep networks and boosting'. In: *Bioinformatics (Oxford, England)* 28.23 (2012), pp. 3066–3072.

P. Fariselli et al. 'Prediction of contact maps with neural networks and correlated mutations'. In: *Protein Engineering* 14.11 (2001), pp. 835–843.

Referências III

P. Fariselli and R. Casadio. 'A neural network based predictor of residue contacts in proteins'. In: *Protein Engineering* 12.1 (1999), pp. 467–475.

Yaping Fang Yunqi Li and Jianwen Fang. 'Predicting residue-residue contacts using random forest models'. In: *Bioinformatics* 27.24 (2011), pp. 3379–3384.

Obrigado